**Chronic Kidney Disease**

**Group – 102**

Reena Nagrale - [2021fa04035@wilp.bits-pilani.ac.in](mailto:2021fa04035@wilp.bits-pilani.ac.in)

Ankur Gupta - [2021fa04040@wilp.bits-pilani.ac.in](mailto:2021fa04040@wilp.bits-pilani.ac.in)

Pratiti Shah - [2021fa04031@wilp.bits-pilani.ac.in](mailto:2021fa04031@wilp.bits-pilani.ac.in)

**Contribution - Equal contribution from each team member.**

**Problem Statement:**

We are provided with csv file with 400 records 25 features. Features are related to medical tests parameter that are observed in people leading to chronic kidney disease. Our goal is to understand and analyse the given data, perform Exploratory Data Analysis and build an Machine Learning model to predict if a person is suffering from CKD for the given medical tests.

**Introduction:**

Chronic kidney disease (CKD) means **your kidneys are damaged and can't filter blood the way they should**. The kidneys filter waste and excess fluid from the blood. As kidneys fail, waste builds up.

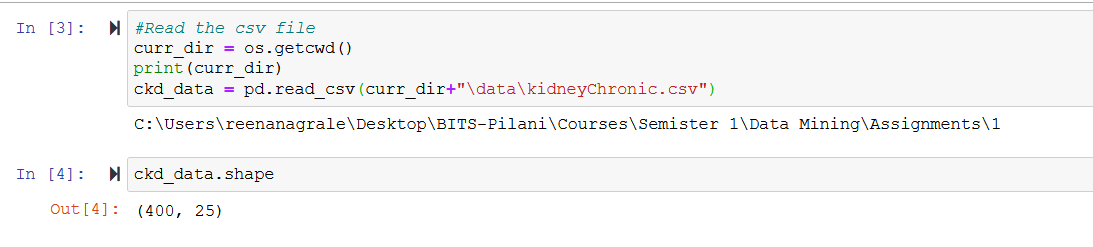
Symptoms develop slowly and aren't specific to the disease. Some people have no symptoms at all and are diagnosed by a lab test.

Medication helps manage symptoms. In later stages, filtering the blood with a machine (dialysis) or a transplant may be required.

The main risk factors for developing kidney disease are diabetes, high blood pressure, heart disease, and a family history of kidney failure.

**Exploratory Data Analysis**

We use python Jupyter notebook to perform analysis on csv data. We are provided with csv file, its imported in Jupyter notebook and stored as dataframe.



Data Pre-processing:

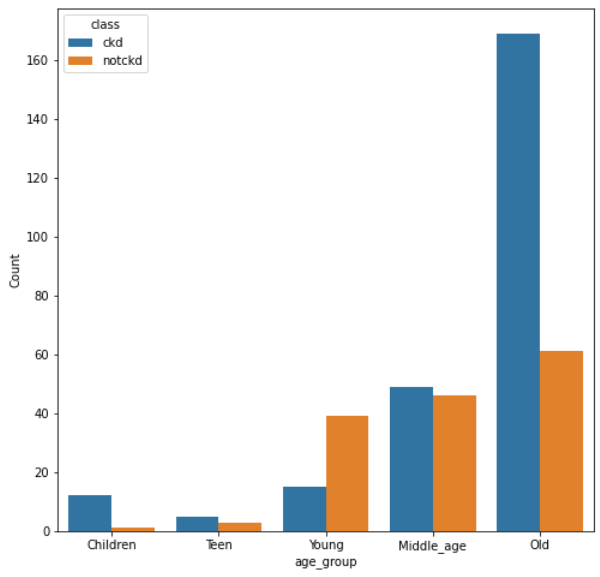
**Steps performed for Data Cleansing:**

1. Evaluate missing or null data
   1. We observe missing\_values as [‘?’, ‘\t’] and no blank records.
   2. Operations on data are performed to fill these values with appropriate records.
2. Understand the features based on the domain knowledge and use the appropriate values to generate new features from available one.
3. Fill the wrong values of some features referencing from related columns. Like diabetes\_mellitus and be obtained from blood\_sugar\_random. So, such techniques are incorporated during data cleansing process.

Data Analysis and some Observations from given dataset

We sliced the data on Age and observed that, major chronic kidney disease is observed for OLD Age people ie 50 and years above.

Middle age people have 50% changes of impacting with chronic kidney disease. We focus to understand what causes more impact of Middle age and Old age people with Kidney diseases.



**Specific gravity**

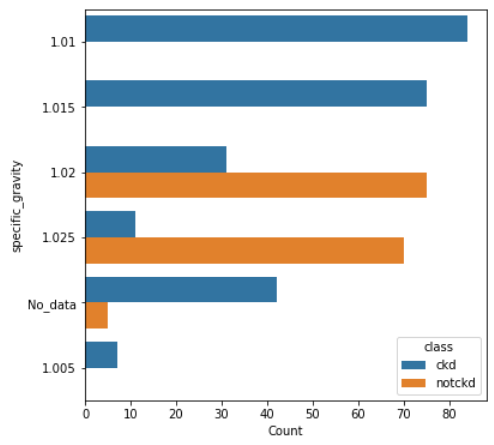
Specific gravity refers to values in range as below and suggests the hydration level index.

well hydrated < 1.010

Minimal dehydrated 1.010 - 1.020

significant dehydrated 1.021 - 1.030

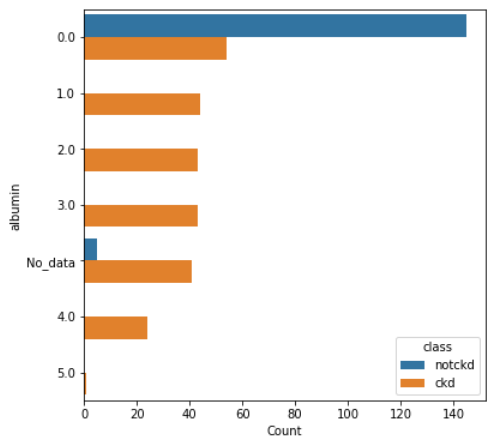
Serious dehydration > 1.030



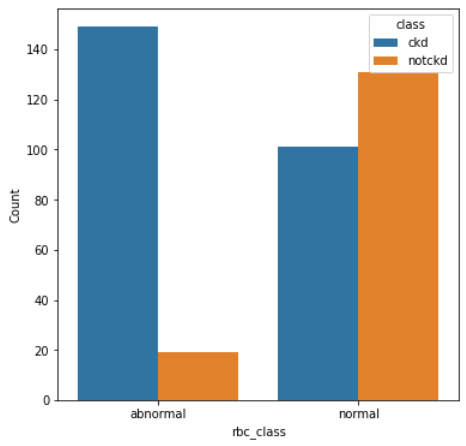
From the graph its suggested that anyone reporting minimal to significantly serious dehydration issues have suffered with chronic kidney disease. Approx. 50% of people have reported with dehydration issues have reported CKD.

**Albuminuria**

Albuminuria is a sign of kidney disease and means that you have too much albumin in your urine. A healthy kidney doesn’t let albumin pass into the urine. A damaged kidney lets some albumin pass into the urine. 0 - Normal, 1-Mild, 2-High,3-Very High, 4-Servere 5-Extreme



It is suggested that anything above normal range of albumin value means person is suffering from CKD.



Normal range of rbcc is 4.2 to 6.1. Anything above or below this range is treated abnormal. So, creating new class as rbc\_class. Based on the observation, it is suggested that people suffering with ckd has reported with abnormal RBC values.

**Data Preparation for Model Building**

**Feature Selection:**

We tested the correlation coefficient of each features using Chi-square test and Point Biserial Correlation Analysis to understand the features that high correlation coefficient for prediction of target.

Based on the Chi-square test we observe that bacteria, cad, pus\_cell\_count column has very high p-value and so we can remove this feature from our dataset and understand the model performance.

Based on the coefficient correlation matrix, we understand that hemoglobin, packed\_cell\_vol, red\_blood\_cell\_count, hypertension,red\_blood\_cell, specific\_gravity, blood\_urea, albumin have high correlation with the target label. These form strong predictors in our model. This is also confirmed with Chi-square test

We performed data transformation using label Encoding and applied Scaling techniques to normalise the data.

**Model Building**

We built and tested different classification models like – LogisticRegression, Decision Tree, RandomForest classifier, SVM and KNN Classifier for 3 conditions.

Part 1.1 - Select the train test data sample by Shuffle = True

Part I.2 - train\_test\_split with shuffle = False

Part I.3 - Model performance with scaled data

We identified, RandomForest classifier, Decision Tree and LogisticRegression give better performance in terms of model accuracy.

All the detailed analysis is part of the Jupyter notebook attached. It can be loaded in Jupyter and executed directly. Placing the csv data file in current working directory/data folder.